**Cach memory**

• Cache memory is a fast small memory where the active portion of the program and data are placed in, so the average memory access time is reduced.

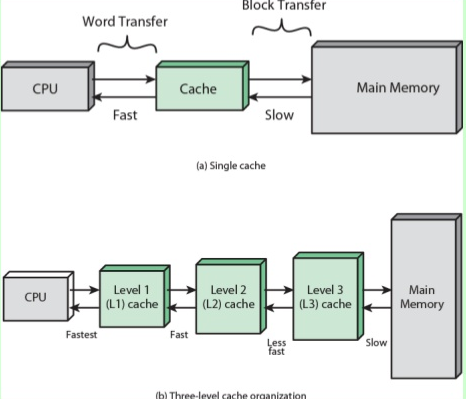
• Thus reducing the total execution time of the program.

•Cache memory is in between CPU and main memory.

• CPU access data from cache.

• L-1 cache fabricated on CPU chip( on chip).

•L-2 cache btween main memory and CPU( off chip).



•When CPU needs to access a memory, cache is examined

• If memory location found then OK! Read ( Hit)

•If not, then memory location is searched in main memory and block that contained the required location is transffered to the cahe and read by CPU (Miss).

**Principle of Locality:**

Programs tends to reuse data and instructions near those they heve used recently. There are two types of locality:

•Temporal locality: Recently referenced items are likely to be referenced in the near future.

•Spatial locality: a neighbor of a recently referenced memory location is likely to be refrenced.

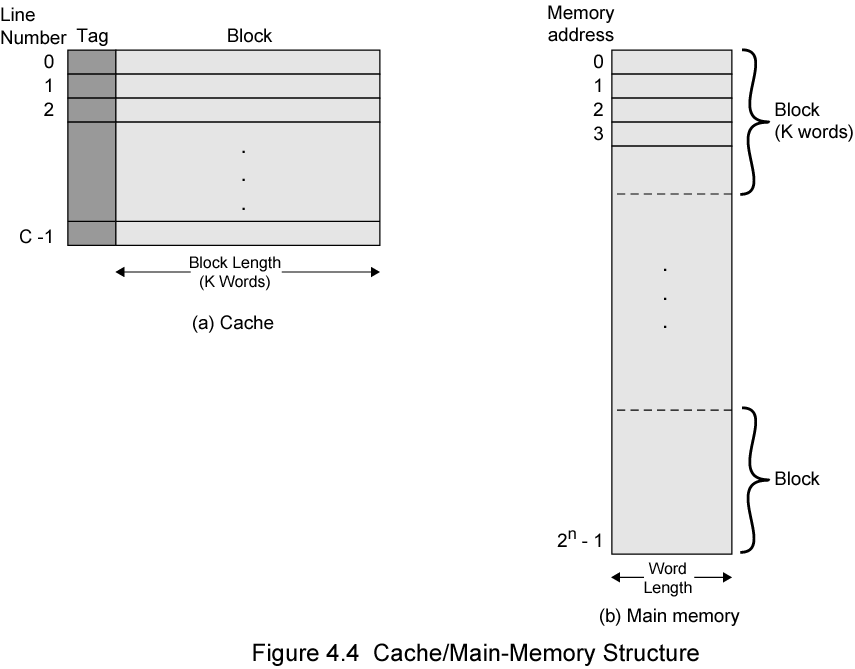
**Cache performance:**

• The performance of cache memory is measured by Hit Ratio.

• Hit Ratio=(total hit)/(total hit +total miss).

• Hit Ratio of 0.9 and higher have been reported.

**Cache/Main Memory Structure**



### Basic elements of cache design:

### • Size

### • Mapping function

### • Replacement algorithm

### • Write policy

### • Block size

### • Number of caches